Bayesian Manifold Learning: the Locally Linear Latent Variable Model (LL-LVM)

Mijung Park[1], Wittawat Jitkrittum[1], Ahmad Qamar[2], Zoltán Szabó[1], Lars Buesing[3], Maneesh Sahani[1]

1: Gatsby Computational Neuroscience Unit, University College London, 2: Thread Genius, 3: Google Deepmind

Manifold Learning

- Learning in high-dim. space is hard and expensive.
- Good news: intrinsic dimensionality is often low.
  - Observations lie on a low-dim. manifold embedded in a high-dim. space.
- **Manifold learning:** uncover the low-dim. manifold structure.

Our Goal

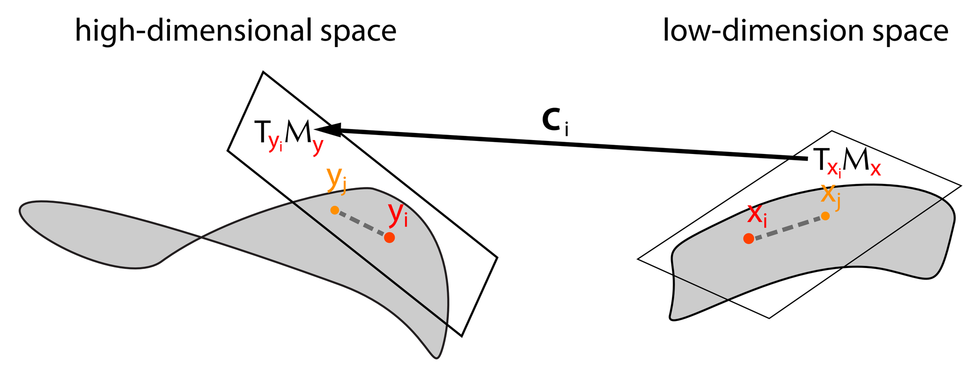
Recover data manifold in a Bayesian probabilistic way, while preserving geometric properties of local neighbourhoods.

Advantages:

- Fully probabilistic. Uncertainty estimates available.
- Principled way to evaluate manifold dimensionality.
- Learned model can handle unseen data points naturally.

Our Approach: LL-LVM

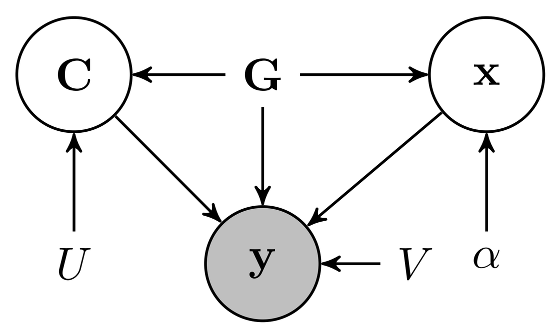
- Assume a *locally linear* mapping between tangent spaces in low and high dimensional spaces



$$\mathbf{y}_j - \mathbf{y}_i \approx \mathbf{C}_i(\mathbf{x}_j - \mathbf{x}_i)$$

- **Input:** neighbourhood graph $\mathbf{G} = [\eta_{ij}]$ with binary adjacency indicator $\eta_{ij} = 1$ if points i, j are neighbours.
- Find posterior distribution $p(\mathbf{C}, \mathbf{x}|\mathbf{y}, \mathbf{G})$ over the linear maps $\mathbf{C} = [\mathbf{C}_1, \cdots, \mathbf{C}_n]$ and the latent variables $\mathbf{x} = [\mathbf{x}_1^\top, \cdots, \mathbf{x}_n^\top]^\top \in \mathbb{R}^{nd_x}$.

Model



Joint distribution:

$$p(\mathbf{y}, \mathbf{C}, \mathbf{x}|\mathbf{G}) = p(\mathbf{y}|\mathbf{C}, \mathbf{x}, \mathbf{G})p(\mathbf{C}|\mathbf{G})p(\mathbf{x}|\mathbf{G}).$$

- **Prior on latent** $\mathbf{x}$: assume neighbouring points are similar,

$$p(\mathbf{x}|\mathbf{G}, \alpha) = \mathcal{N}(\mathbf{0}, \mathbf{\Pi}) \propto -\tfrac{1}{2}\sum_{i=1}^{n}\left(\alpha||\mathbf{x}_i||^2 + \sum_{j=1}^{n}\eta_{ij}||\mathbf{x}_i - \mathbf{x}_j||^2\right),$$

where α controls the expected scale, $\mathbf{\Pi}^{-1} = \alpha\mathbf{I}_{nd_x} + \mathbf{\Omega}^{-1}$, $\mathbf{\Omega}^{-1} = 2\mathbf{L}\otimes\mathbf{I}_{d_x}$ and $\mathbf{L} = \mathrm{diag}(\mathbf{G1}) - \mathbf{G}$.

- **Prior on linear maps**: matrix normal,

$$p(\mathbf{C}|\mathbf{G}, \mathbf{U}) = \mathcal{MN}(\mathbf{0}, \mathbf{U}, \mathbf{\Omega}), \quad \text{where } \mathbb{E}[\mathbf{CC}^\top] \propto \mathbf{U},\ \mathbb{E}[\mathbf{C}^\top\mathbf{C}] \propto \mathbf{G}.$$

- **Likelihood**: penalise the approximation error,

$$p(\mathbf{y}|\mathbf{C}, \mathbf{x}, \mathbf{V}, \mathbf{G}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{y}, \mathbf{\Sigma}_\mathbf{y})$$

$$\propto -\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\eta_{ij}((\mathbf{y}_j - \mathbf{y}_i) - \mathbf{C}_i(\mathbf{x}_j - \mathbf{x}_i))^\top\mathbf{V}^{-1}((\mathbf{y}_j - \mathbf{y}_i) - \mathbf{C}_i(\mathbf{x}_j - \mathbf{x}_i)),$$

where $\mathbf{V}^{-1} = \gamma\mathbf{I}$ and γ is to be learned.

Variational EM

- Maximising log marginal likelihood is intractable. Maximise lower bound $\mathcal{F}$ instead

$$\log p(\mathbf{y}|\mathbf{G}, \boldsymbol{\theta}) \geq \iint q(\mathbf{C}, \mathbf{x})\ \log\frac{p(\mathbf{y}, \mathbf{C}, \mathbf{x}|\mathbf{G}, \boldsymbol{\theta})}{q(\mathbf{C}, \mathbf{x})}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{C} := \mathcal{F}(q(\mathbf{C}, \mathbf{x}), \boldsymbol{\theta}).$$

- For computational tractability, assume $q(\mathbf{C}, \mathbf{x}) = q(\mathbf{x})q(\mathbf{C})$.
- Variational expectation maximisation (EM) algorithm:
  - E-step for computing $q(\mathbf{C}, \mathbf{x})$ by

$$q(\mathbf{x}) \propto \exp\left[\int q(\mathbf{C})\log p(\mathbf{y}, \mathbf{C}, \mathbf{x}|\mathbf{G}, \boldsymbol{\theta})\mathrm{d}\mathbf{C}\right] = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x}),$$

$$q(\mathbf{C}) \propto \exp\left[\int q(\mathbf{x})\log p(\mathbf{y}, \mathbf{C}, \mathbf{x}|\mathbf{G}, \boldsymbol{\theta})\mathrm{d}\mathbf{x}\right] = \mathcal{N}(\mathbf{c}|\boldsymbol{\mu}_\mathbf{c}, \mathbf{\Sigma}_\mathbf{c}).$$

  - M-step for learning $\boldsymbol{\theta} = \{\alpha, \mathbf{U}, \gamma\}$,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{F}(q(\mathbf{C}, \mathbf{x}), \boldsymbol{\theta}).$$

Illustration 1: Mitigating Short-Circuiting Problems

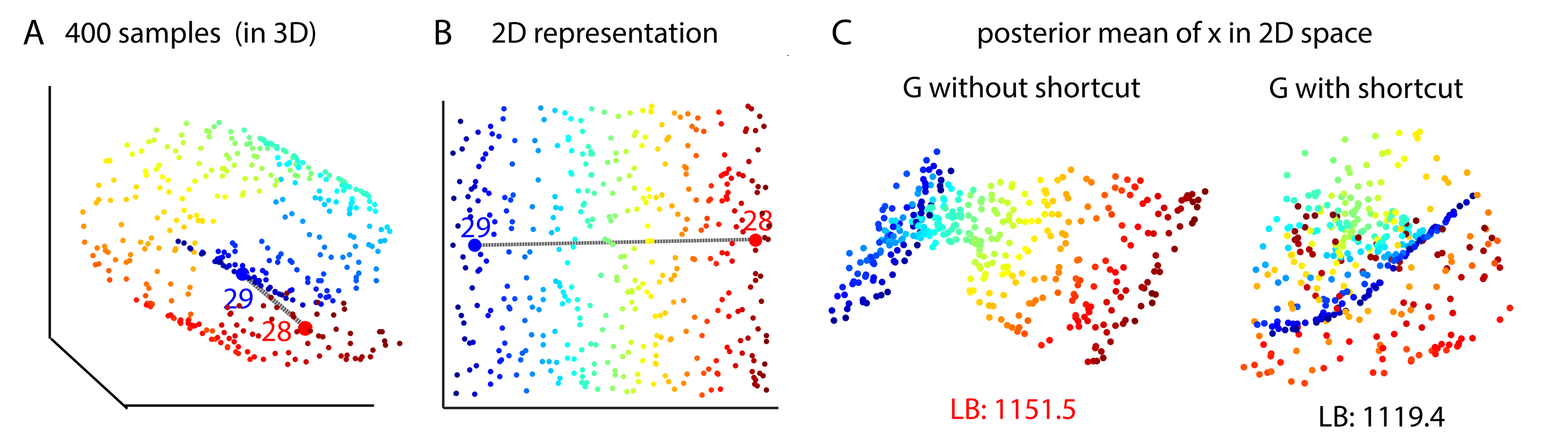


Figure : (**A**) Two datapoints seem close to each other, (**B**) but actually far in 2D space. (**C**) Short-circuiting the two datapoints lower the lower bound.

- The lower bound $\mathcal{F}$ can be used to evaluate a hypothesised neighbourhood structure.

Illustration 2: Modelling USPS Handwritten Digits

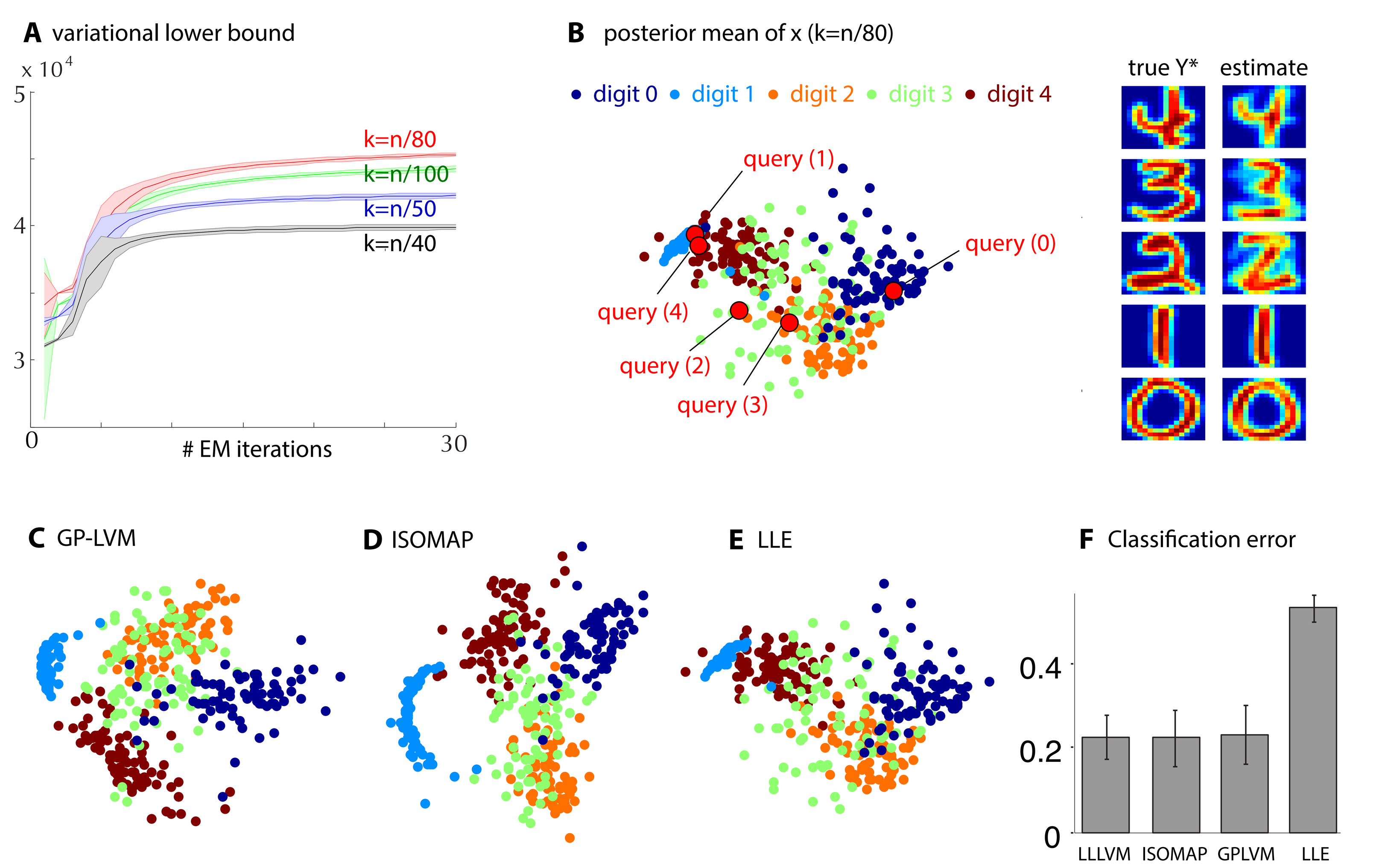
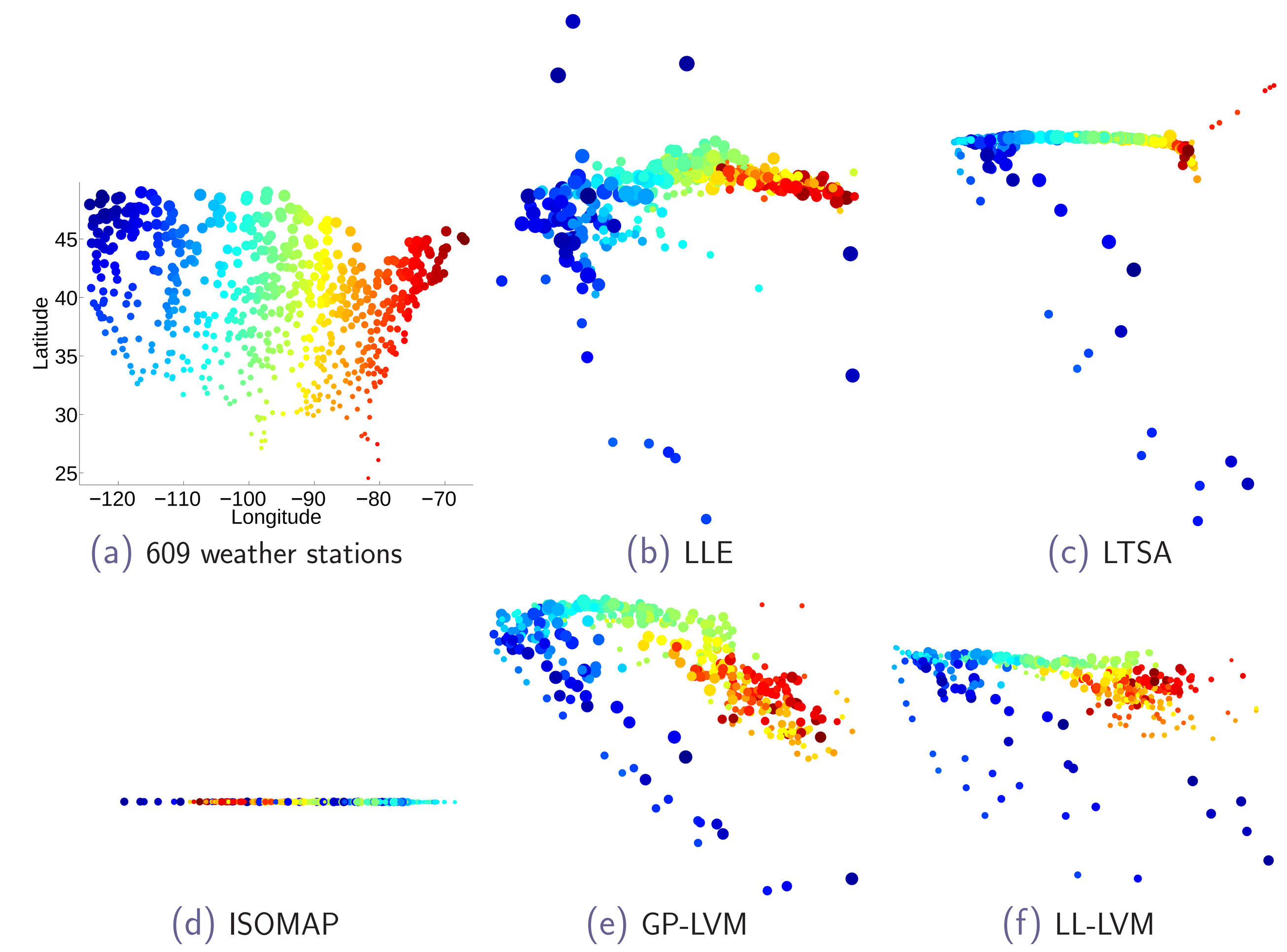


Figure : (**A**): Variational lower bound with different k's (#neighbours). (**B**): Posterior mean of $\mathbf{x}$ by LL-LVM. (**F**): 1-NN classification error on test data using the inferred $\mathbf{x}$.

- Classification with LL-LVM coordinates outperforms GP-LVM and LLE, and matches ISOMAP.

Illustration 3: Mapping Climate Data

- **Goal:** Recover 2D geographical relationships between weather stations.
- $\mathbf{y}_i$ = 12-dim. vector of monthly precipitation measurements at a weather station.



- The projection obtained from LL-LVM recovers the topological arrangement of the stations to a large degree.

Gaussian Process Latent Variable Model (GP-LVM)[1, 2]

- Define a mapping from latent $\mathbf{X}$ to data $\mathbf{Y}$ using GP.
- For data $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{d_y}] \in \mathbb{R}^{n\times d_y}$ and latents $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{d_x}] \in \mathbb{R}^{n\times d_x}$,

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{k=1}^{d_y}\mathcal{N}(\mathbf{y}_k|\mathbf{0}, \mathbf{K} + \beta^{-1}\mathbf{I}_n),$$

where the i, jth element of the covariance matrix is $k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2\exp\left[-\frac{1}{2}\sum_{q=1}^{d_x}\alpha_q(x_{i,q} - x_{j,q})^2\right]$, and α_q's determine dimensionality of latent space.

- **Limitations**:
  - No preservation of local neighbourhood properties
  - Smoothness of manifold constrained by pre-chosen covariance function.
  - Use auxiliary variable for variational inference. Restrict the choice of covariance function.

Relationship of LL-LVM and GP-LVM

Integrating out $\mathbf{C}$ from likelihood yields

$$p(\mathbf{y}|\mathbf{x}, \mathbf{G}, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{C}, \mathbf{x}, \mathbf{G}, \boldsymbol{\theta})p(\mathbf{C}|\mathbf{G}, \boldsymbol{\theta})d\mathbf{C} = \frac{1}{Z_{Y_y}}\exp\left[-\frac{1}{2}\mathbf{y}^\top\mathbf{K}_{LL}^{-1}\mathbf{y}\right].$$

- In contrast to GP-LVM, the precision matrix $\mathbf{K}_{LL}^{-1}$ is *directly* determined by the *graph structure* given the observations.

$$\mathbf{K}_{LL}^{-1} = (2\mathbf{L}\otimes\mathbf{V}^{-1}) - (\mathbf{W}\otimes\mathbf{V}^{-1})\,\mathbf{\Lambda}\,(\mathbf{W}^\top\otimes\mathbf{V}^{-1}),$$

where $\mathbf{W}$ is a function in $\mathbf{x}$ and $\mathbf{L}$ and $\mathbf{\Lambda}$ is a function in $\mathbf{x}^\top\mathbf{x}$ and $\mathbf{L}$.

Conclusion

A new probabilistic approach to manifold learning preserving local geometries in data and equipped with straightforward variational inference.

References

[1] N.D. Lawrence. GP-LVM. *NIPS 2003*.
[2] M.K. Titsias, N.D. Lawrence. Bayesian GP-LVM. *AISTATS*, 2010.

This work was funded by the Gatsby Charitable Foundation.

Contact: mijung@gatsby.ucl.ac.uk